

# Using Apollo at the i5k Workspace@NAL

Monica Poelchau, USDA-ARS NAL  
AGS 2018 Bioinformatics Workshop  
June 7<sup>th</sup>, 2018

PLEASE fill out the post-workshop survey!

<https://tinyurl.com/ybppr8pq>

# Agenda

- Part 1:
  - Manual annotation general overview
  - I5k Workspace tools for manual annotation
    - BLAST, Clustal, HMMER
    - Apollo
  - Manual annotation example: preparation
  - Manual annotation live example
- Part 2:
  - Hands-on exercises

# Other resources

- Monica Munoz-Torres from the Apollo group has a number of comprehensive tutorials:
  - <https://www.slideshare.net/MonicaMunozTorres/presentations>
    - I recommend these slides if you need more background:
      - <https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015>
    - Note - there are two versions of Apollo. Some organisms at the i5k Workspace still use the older version with a slightly different interface
  - If you are new to Apollo, or need a refresher, I **highly recommend** that you review one of her presentations
- The official Apollo annotation guide:
  - <http://genomearchitect.org/users-guide/>
- Other manual curation tutorials:
  - <https://i5k.nal.usda.gov/manual-curation-example>
  - <http://genomecuration.github.io/genometrain/d-feature-curation-crossing/>

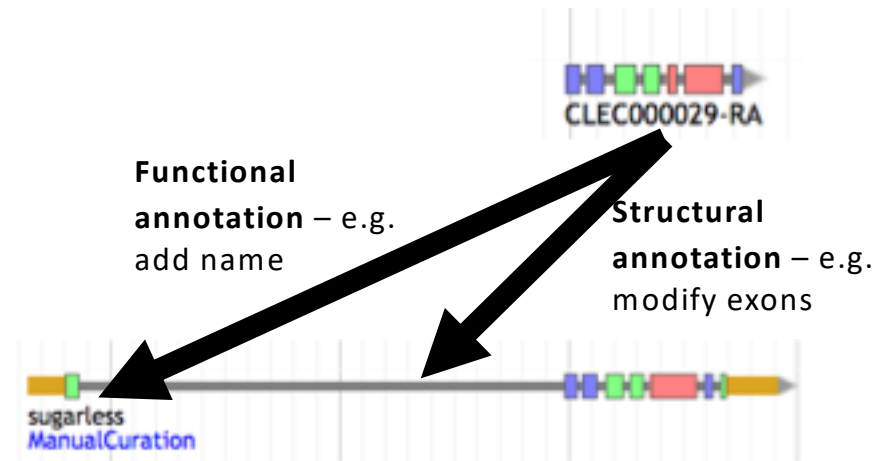
# Quick survey – why do you want to learn how to use Apollo?

1. I work on a gene family that is poorly predicted by annotation pipelines.
2. I work on a particular set of genes and need to verify their predicted structure.
3. I don't want to learn Apollo, but I need to manually annotate in order to get my genome assembly paper published.
4. Just curious.

# Manual annotation general overview

# What is manual annotation?

- Manual review and improvement of an existing gene prediction
- Draw on external evidence (e.g. RNA-Seq, cDNA, genes from other species) to improve a computationally predicted gene model



# Why manually annotate?

- “Incorrect annotations poison every experiment that makes use of them ... Worse still, the poison spreads because incorrect annotations from one organism are often unknowingly used by other projects to help annotate their own genomes.”
  - Yandell and Ence 2012, doi:10.1038/nrg3174
- Link gene models to existing literature and ontologies, providing richer data

# General process of manual annotation

1. Select a chromosomal region of interest (e.g. scaffold)
  1. E.g. find sequence of interest from one or several other species, and align against proteins or genome sequence from your species
2. Select appropriate evidence (tracks in Apollo, or your own files)
3. Determine whether a feature in your evidence provides a reasonable starting gene model
  1. If yes: select and drag the feature to the 'user-created annotations' area, creating an initial gene model. If necessary use editing functions to adjust the model.
  2. If not – get in touch with us!
4. Edit model if necessary
5. Check your edited gene model for integrity and accuracy by comparing it with available homologs
  1. Verify that the gene model is the best representation of the underlying biology
6. Repeat steps 1 through 5 as needed to refine model
7. Add annotation details in the "Information Editor"
  1. Name, symbol, other comments

Adapted from <https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015>

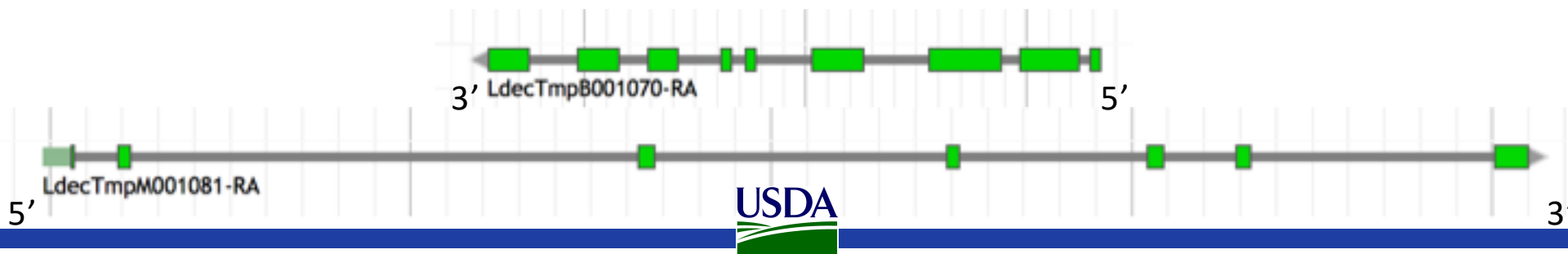
# I5k Workspace 'Etiquette'

1. Use Apollo to improve a gene model in an i5k Workspace assembly.
  1. If you just want to practice – use one of our training instances.
    1. <https://i5k.nal.usda.gov/browseapollo-training>
  2. If you just want to view the data – you probably can get what you want without using Apollo. All of the data that we host is public.
2. Your annotation work is a community effort.
  1. If you notice that someone else is working on your model of choice, get in touch with them (or us) and collaborate – don't make a 2nd model or delete the other model.
  2. Keep in mind that your work may be used by the scientific community once you're done.
3. If you publish any of your work generated in the i5k workspace:
  1. Get in touch with the genome contact first (you can find the contact info on the organism page; <https://i5k.nal.usda.gov/species>);
  2. Please cite the i5k Workspace paper! This helps us continue to exist.
    1. <https://doi.org/10.1093/nar/gku983>

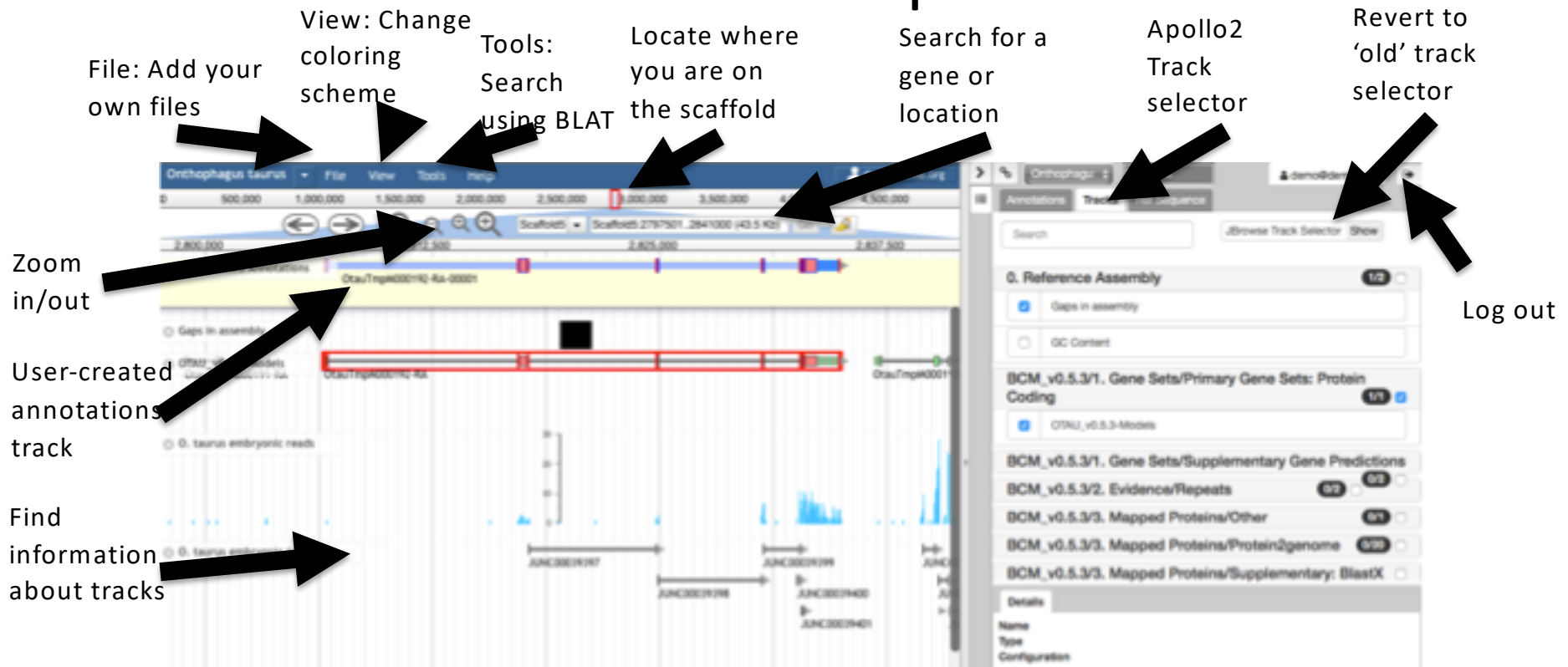
# Manual annotation: i5k Workspace tools

# First, some conventions

- HSP – High scoring pair in BLAST/BLAT alignments
  - The ‘Hits’ in an alignment result set
  - A subsection of a pair of sequences with sufficient score
  - HSPs can change based on the alignment parameters
- Five prime end and three prime end
  - Based on direction of transcription
  - Initiation site is at the five prime end
  - Stop codon is at the three prime end
- In the genome browser, arrowheads indicate direction



# JBrowse and Apollo2



JBrowse is a web- based genome browser    Apollo adds editing functions to JBrowse

- Visualize features that are mapped to a genome
- These features are displayed as tracks
- Many different types of data may be displayed

- Manual gene curation
- Changes automatically saved back to server
- Edits are visible to other annotators in real-time
- Editing history is tracked

# Apollo2 – Annotations Panel

The screenshot shows the Apollo2 web interface for genome annotations. The main panel displays a genomic track with various annotations, including a user-created 'test mRNA' and several junctions (JUNC00039387, JUNC00039388, JUNC00039389, JUNC00039390). The right-hand 'Annotations panel' allows for filtering and viewing details of specific annotations. The 'Filter annotations' section includes fields for 'Annotation Name', 'Reference Sequence', 'All Types', and 'All Users'. Below this is a table of annotations with columns for Name, Seq, Type, Length, and Updated. The 'test mRNA' annotation is highlighted. The 'Details' tab for 'test mRNA' shows its description, location, reference sequence, and owner. The 'Coding' tab provides a detailed view of the mRNA's structure, including a table of exons and introns, and controls for modifying individual features like the 5' End, 3' End, and Strand.

Annotations panel

Filter annotations

View annotation overview

Click on arrow to jump to annotation

View functional annotation details

Can modify individual features via 'Coding' Panel

Type	Start	Length
exon	2,817,179	677
exon	2,806,566	226
exon	2,832,890	2,052
CDS	2,806,608	27,094
exon	2,824,806	152
exon	2,832,682	150
exon	2,830,616	128

# Apollo2 – Ref Sequence Panel

The screenshot displays the Onthophagus taurus genome browser. The top panel shows the genome assembly with a scale from 0 to 3,000,000. Below this, a track labeled 'Scaffold1' shows a specific region (455,000 to 460,000) with 'User-created Annotations' and 'OtauTimpA001262-RA-00001'. The bottom panel shows 'Gaps in assembly', 'OTAU\_v0.5.3-Models', 'O. taurus embryonic reads', and 'O. taurus embryonic junctions'. On the right, a 'Ref Sequences' panel is open, showing a list of scaffolds (Scaffold5, Scaffold54, Scaffold527, etc.) with their lengths and annotations. Arrows point to various features: 'Reference sequence panel', 'Filter sequences', 'Export sequences/annotation gff3', and 'View reference sequence list'.

Export 1 sequence(s) from *Orthopagus taurus* as GFF3

☒ GFF3 ☐ GFF3 with FASTA

**Export** Cancel

Name	Length	Annotations
Scaf001	4,952,830	
Scaf004	1,101,439	
Scaf027	1,104,991	
Scaf042	996,189	

# i5k Workspace BLAST: one way to access Apollo

The screenshot shows the i5k Workspace BLAST web application. The interface includes a top navigation bar with 'i5k@NAL', 'Tools', 'About Us', and 'Contact'. The main content area is titled 'BLAST Databases' and contains several sections:

- Organisms:** A list of organisms with radio buttons. *Eurytemora affinis* is selected.
- Eurytemora affinis:** A sub-section with radio buttons for 'Nucleotide' (selected), 'Genome Assembly' (with a file path), 'Transcript', and 'Peptide'.
- Query Sequence:** A text area containing a peptide sequence: `>FBpp0070332  
MDNCDQDAFRRLSHKEEVKPDISQNDNN  
SGSPKAEIPNPFM QAMGMVHVLPGSNGASS  
NNNSAGDAQMAQAPNSAG  
GSAAAVQOQYPPNHPLSGSKHLCSICGDRA  
SGQHYGVYSCEGCKGFFKRTVRKDLTYACRE`. Below it is a 'Browse...' button and the text 'No file selected.'.
- Program:** Radio buttons for 'blastn', 'blastx', 'tblastn', 'tblastx', 'blastp', and 'blastf'. 'tblastn' is selected.

Annotations with arrows point to specific elements:

- 'Select organism' points to the *Eurytemora affinis* selection in the Organisms list.
- 'Paste or upload query sequence(s)' points to the Query Sequence text area.
- 'Program is automatically selected' points to the 'tblastn' radio button.
- 'Select organism-specific database' points to the 'Genome Assembly' radio button.
- 'BLAST against the genome assembly to view HSPs in Jbrowse' points to the 'Genome Assembly' radio button.

URL: <https://i5k.nal.usda.gov/webapp/blast/>

# i5k Workspace BLAST: one way to access Apollo

BLAST result page with 4 panels

Click on blue blastdb icon next to your favorite HSP

Blast results are displayed in Apollo

# HMMER and Clustal

- Use HMMER to detect remote protein homologs
- <https://i5k.nal.usda.gov/webapp/hmmer/>
- Use Clustal to perform multiple sequence alignments
- <https://i5k.nal.usda.gov/webapp/clustal/>

# Tips and Tricks

- The i5k Workspace BLAST results persist for one week
  - You can bookmark and share searches
  - BLAST HSPs are ‘draggable’ and can be used in annotations
- Jbrowse/Apollo URLs can be shared
  - Allow you to share the exact view (including active tracks) with others
  - Great for troubleshooting with collaborators
- In Apollo “walk” feature boundaries
  - Square brackets walk exon boundaries: [ and ]
  - Curly brackets walk gene boundaries: { and }
- In Apollo, you can pin tracks to the top
- If you know the name or ID of the gene that you’d like to annotate, you can paste it into the search box in Apollo to navigate to it

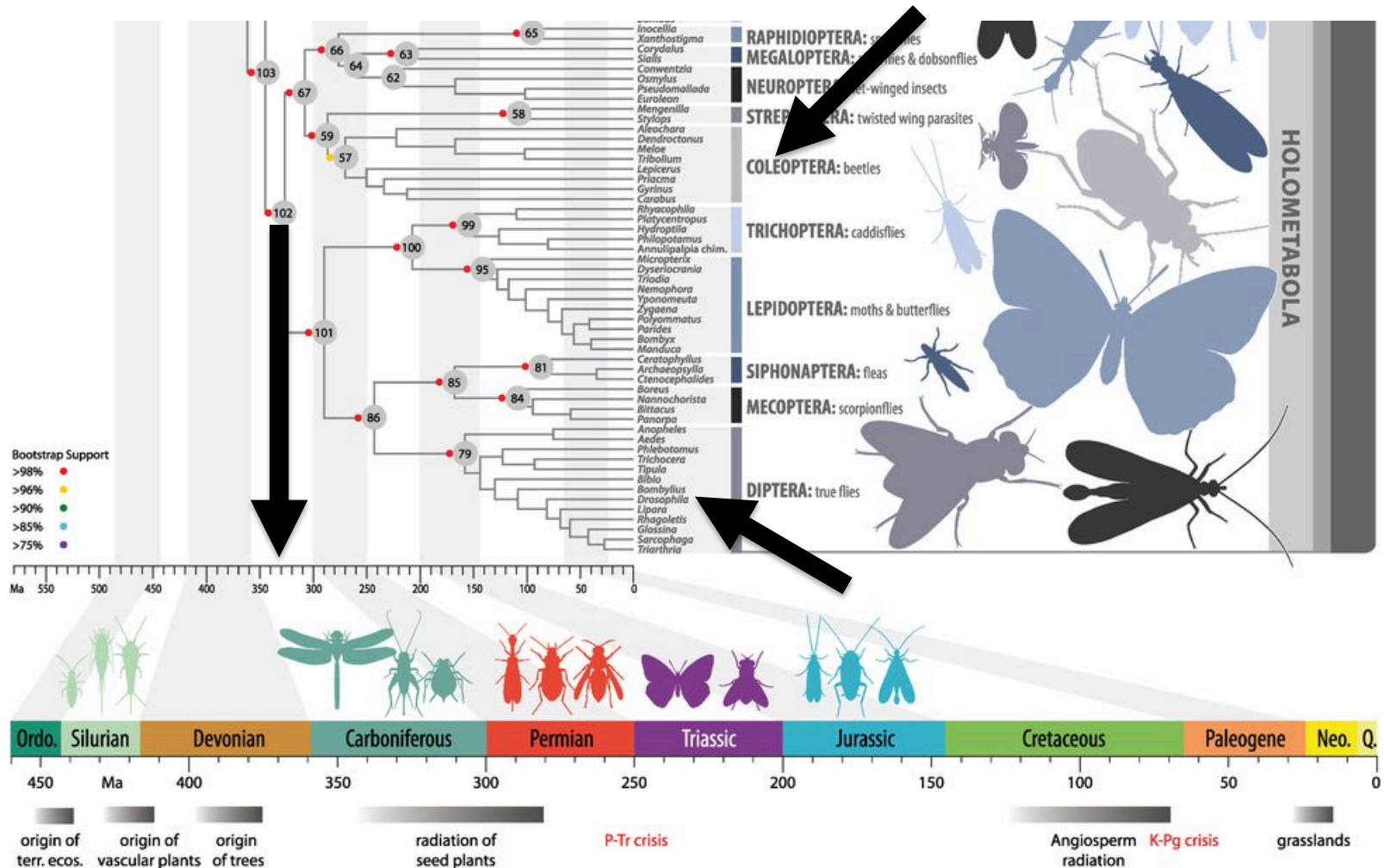
# Manual annotation example: preparation

# Annotation Example

- Alpha-catenin in the beetle *Onthophagus taurus*
  - “Associates with the cytoplasmic domain of a variety of cadherins”.

(<http://www.uniprot.org/uniprot/P35220>)
- More information about the *O. taurus* genome project:  
[https://i5k.nal.usda.gov/Onthophagus\\_taurus](https://i5k.nal.usda.gov/Onthophagus_taurus)

# Notes on *O. taurus* genome



Excerpt of Figure 1 from Misof, Bernhard, et al. "Phylogenomics resolves the timing and pattern of insect evolution." *Science* 346.6210 (2014): 763-767.

# Notes on *O. taurus* genome/browser

- Big advantage for annotation: lots of RNA-Seq and transcriptome data are available to use as contributing evidence for your gene models

# Available tracks for *O. taurus*

**Available Tracks**

☒ filter tracks

- 0. Reference Assembly 2
  - ☐ GC Content
  - ☒ Gaps in assembly
- BCM\_v0.5.3 48
- 1. Gene Sets 3
  - Primary Gene Sets: Protein Coding 1
    - ☒ OTAU\_v0.5.3-Models
  - Supplementary Gene Predictions 2
- 2. Evidence 2
- 3. Mapped Proteins 41
- 4. Transcriptome 2
- Gene Silhouette Models 1
- Transcriptome 13
- Coverage Plots (BigWig) 7
  - ☐ O. taurus F HC L3L+PP1+PD1+ADH
  - ☐ O. taurus F PD1 BRH+CHE+THE+GEN transcripts
  - ☐ O. taurus M HC L3L+PP1+PD1+ADH
  - ☐ O. taurus M PD1 BRH+CHE+THE+GEN transcripts
  - ☒ O. taurus embryonic reads
  - ☐ Onthophagus taurus early female pupa brain transcripts (Coverage Plot)
  - ☐ Onthophagus taurus early male pupa brain transcripts (Coverage Plot)
- Mapped Reads 5
  - ☐ O. taurus embryonic reads
  - ☐ O. taurus whole body, mixed stage RNAseq, reduced coverage, females
  - ☐ O. taurus whole body, mixed stage RNAseq, reduced coverage, males
  - ☐ Onthophagus taurus early female pupa brain transcripts
  - ☐ Onthophagus taurus early male pupa brain transcripts
- Splice Junctions 1
  - ☒ O. taurus embryonic junctions

**Onthophagus taurus**

0 500,000 1,000

← →

User-created Annotations

☐ Gaps in assembly

☐ OTAU\_v0.5.3-Models

☐ O. taurus embryonic r

☐ O. taurus embryonic j

- Gap and GC content tracks
- Baylor Maker annotations:
  - Primary Gene Set:
    - OTAU\_v0.5.3-Models
  - Other tracks that were used to generate the primary gene set
- Additional Gene Silhouette gene predictions
- Transcriptome/RNA-Seq
  - Transcriptome assemblies
  - Coverage plots, Mapped RNA-Seq data, Splice junctions

# Choosing reference proteins: *D. melanogaster* Alpha-cat in UniProt

Annotation score is a heuristic for annotation quality

UniProtKB - P35220 (CTNA\_DROME)

Display

Entry

Publications

Feature viewer

Feature table

BLAST Align Format Add to basket History

Protein | Catenin alpha

Gene | alpha-Cat

Organism | *Drosophila melanogaster* (Fruit fly)

Status | Reviewed - Annotation score: ★★★★★ - Experimental evidence at protein level<sup>1</sup>

Organism-specific databases

FlyBase<sup>i</sup>

FBgn0010215 alpha-Cat

Flybase is another great resource

UniProtKB - P35220 (CTNA\_DROME)

Display

Entry

Publications

Feature viewer

Feature table

BLAST Align Format Add to basket History



▼ Molecule processing

Chain

► Sequence information

► Mutagenesis

► Proteomics

ProtVista<sup>1</sup>



Feature viewer gives graphical view of domains and sites

Associates with the cytoplasmic domain of a variety of cadherins.

Source: <http://www.uniprot.org/uniprot/P35220>

# Choosing reference proteins: *Tribolium castaneum* Alpha-cat

- Find orthogroup at OrthoDB:

**OrthoDB**

UNIVERSITÉ DE GENÈVE Zdobnov's Computational Evolutionary Genomics group

OrthoDB v9.1

The Hierarchical Catalog of Orthologs v9.1

OrthoDB is a comprehensive catalog of orthologs, i.e. genes inherited by extant species from their last common ancestor. Arising from a single ancestral gene, orthologs form the cornerstone for comparative studies and allow for the generation of hypotheses about the inheritance of gene functions. Each phylogenetic clade or subclade of species has a

Build your query Search by sequence

Copy a protein sequence (<1000 a.a.):

hspIP36233CTNA.DROME Canin alpha C5-Drosophila melanogaster OM-7327 SH-alpha-Cat R1-1 Sh-2 MLKPKMSTLPDQALYKWDPANLERTMSYKTLERKVLGV FTLYNTGSPSKKKKGS KRASACAAAEKATENFIQDQAYENPDTQEMLTN GDAMSHAAEPSED CSSLKRGNNVVAAPNLSKYTRLLULADNVDNILL DLNKLKNASSGDELMD NMSQFGRADELKGAHAKRGQELKDPQRDDLAARAKLKH ETALLTKKKKKKKKKKK

Your search returned 14 groups

Bookmark [OrthoDB](#) | [Get All Fasta](#) | [Get All as Tab delimited](#)

Group [EOG090R0115](#) at Endopterygota level 113 genes in 97 species

Group [EOG090W014M](#) at Insecta level 139 genes in 113 species

**Tribolium castaneum**

**TC004609** Putative uncharacterized protein 906 6

Ensembl: [TC004609](#) Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:D6W728]

GO Molecular Function: structural molecule activity; calcium ion binding; actin filament binding

GO Biological Process: cell adhesion

GO Cellular Component: actin cytoskeleton

Entrez: [LOC658703](#)

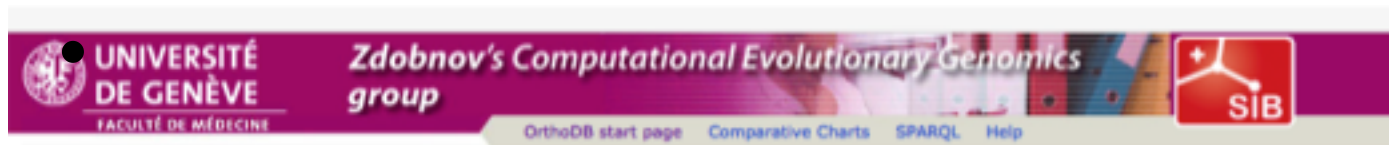
WikiGene: [LOC658703](#)

Paste in sequence from Uniprot

Choose Endopterygota

Find Tribolium gene name

# Choosing reference proteins: *Tribolium castaneum* Alpha-cat



OrthoDB

Your search for **EOG090R0115** returned 1 group

Bookmark [OrthoDB](#) | [Get All Fasta](#) | [Download](#)

[Polypedilum vanderplanki](#)

Gene ID	Protein ID	Length
1	Pv.10391	1640
2	Pv.10390	11265

Get protein sequences for this orthogroup

1 match < >  Done

```
>7070:000d3d {"pub_gene_id":"TC004609", "pub_og_id":"EOG090R0115", "og_name":"actin binding", "level":33392, "description":"Putative uncharacterized protein "}
MKDGFOLKMDPFENLEINTMSVEXTLEPLVLQVTTLVNTRGPKKKKKKSKRANALVSTYKATENTFIEKGEQIAYENPDITQEMLSAVESVRKTCGRN...
FSEDFCSLLKRCNNVRAARNLLSAVTRLLIADNVVDVHLLKSLHVVNDIEKLNASSQGEILLDNIKAFQGNANELMNQAAKQQLKDFQLRDDLAAM...
KKSTHLLTASHVTVVPELAAAKANRDVYLQVCAVNTINDVAQORTTPQATGPTDGPGLAAALDDFDHNVMEPLAYNEVHTRPGLSEERLESIIISGAALN...
ADSSCTREDEKRIIVASCNVAVQALQQLSEYMSNIGNKESESLNRAIDNMGKTRDLRRQLKAVVDHVSDFLETNVPLLVLIKAAQNGNEKEVEEFVAVMF...
TERENKLVETVANLVCSMNSNEDGVKMYRYAAQIDNLCPEVINAARILAAKPRSKVAQENMDAFKQSWENQVRILTEAVDDITTIIDDFLAVSEMHILEDVNEKV...
LALQSGDADTLDRAGGIRGSGNRVCNVVTAEMDNYPECIYTKRVLEAVKVLNDQVMPKPSQRVQVAVQALSSVPTKEVDENDFIDASRLVTDGVREIRRAVLN...
KKADEDLDPEDELVDENYTLTSAKSSAHTGEGVDEYFDISGITTAREAMKMPSEEDKQKILQQVEYFRSEKLKFDREVAKWDGTNDIIVLAKHMCINMEN...
TDFTRGRCPLKTTMDVINAAKKISEAGTKLDELTRQIAEQCPESSTKKDELLAYLQRIALYCHOMNITSKVKADVONISGELIVSGLDSATSLIQAANKLNMAVV...
LTVKASTVASTKYPRQQTISSPIVVMKMAPEKKPLVRPEKPEEVRAKVRKGSQKKVQNPPIHALSEFQSPTESTI
-----
```

Search for gene ID

# Manual annotation live example

# BLAST dmel, tcas proteins against against *O. taurus* genome

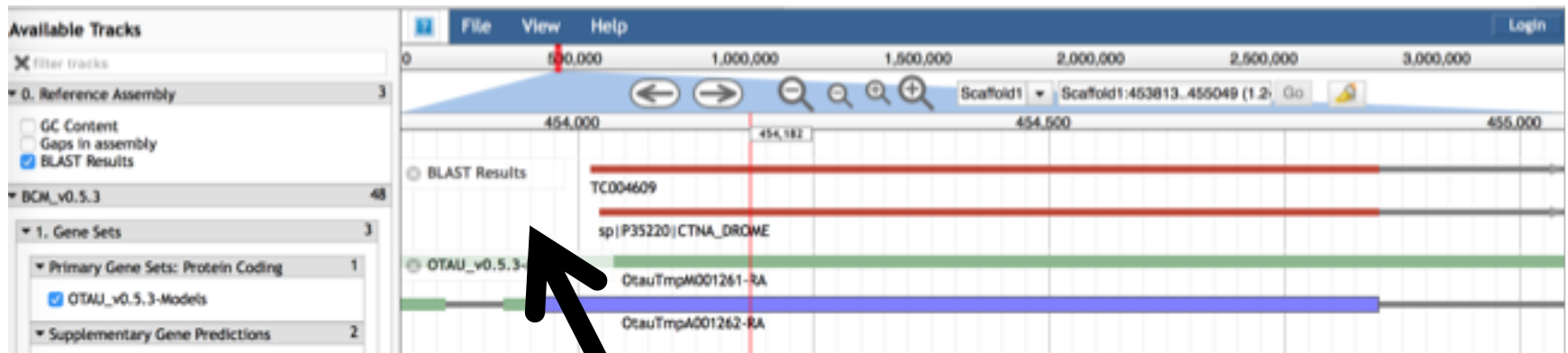
<https://i5k.nal.usda.gov/training/webapp/blast/>



Click on blue blastdb button next to your favorite HSP to view it in JBrowse

Results are filtered by e-value and sorted by position

# BLAST dmel, tcas proteins against against *O. taurus* genome

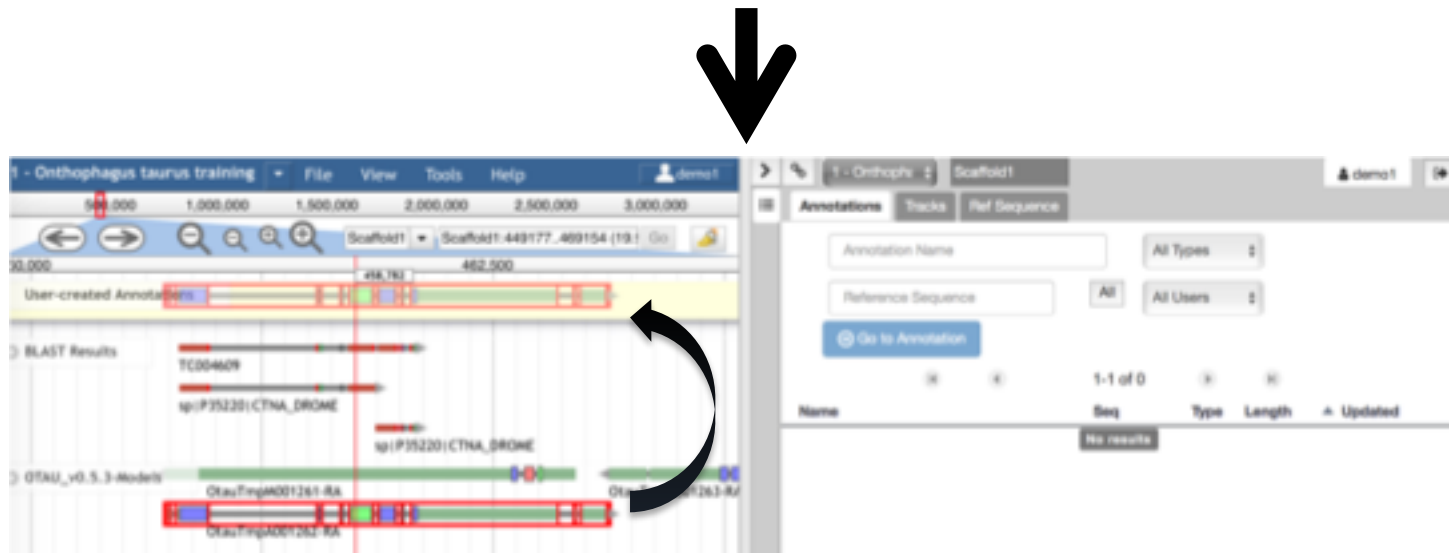


BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality

# Create annotation in user-created annotations track



Log in with  
your  
Apollo  
credentials

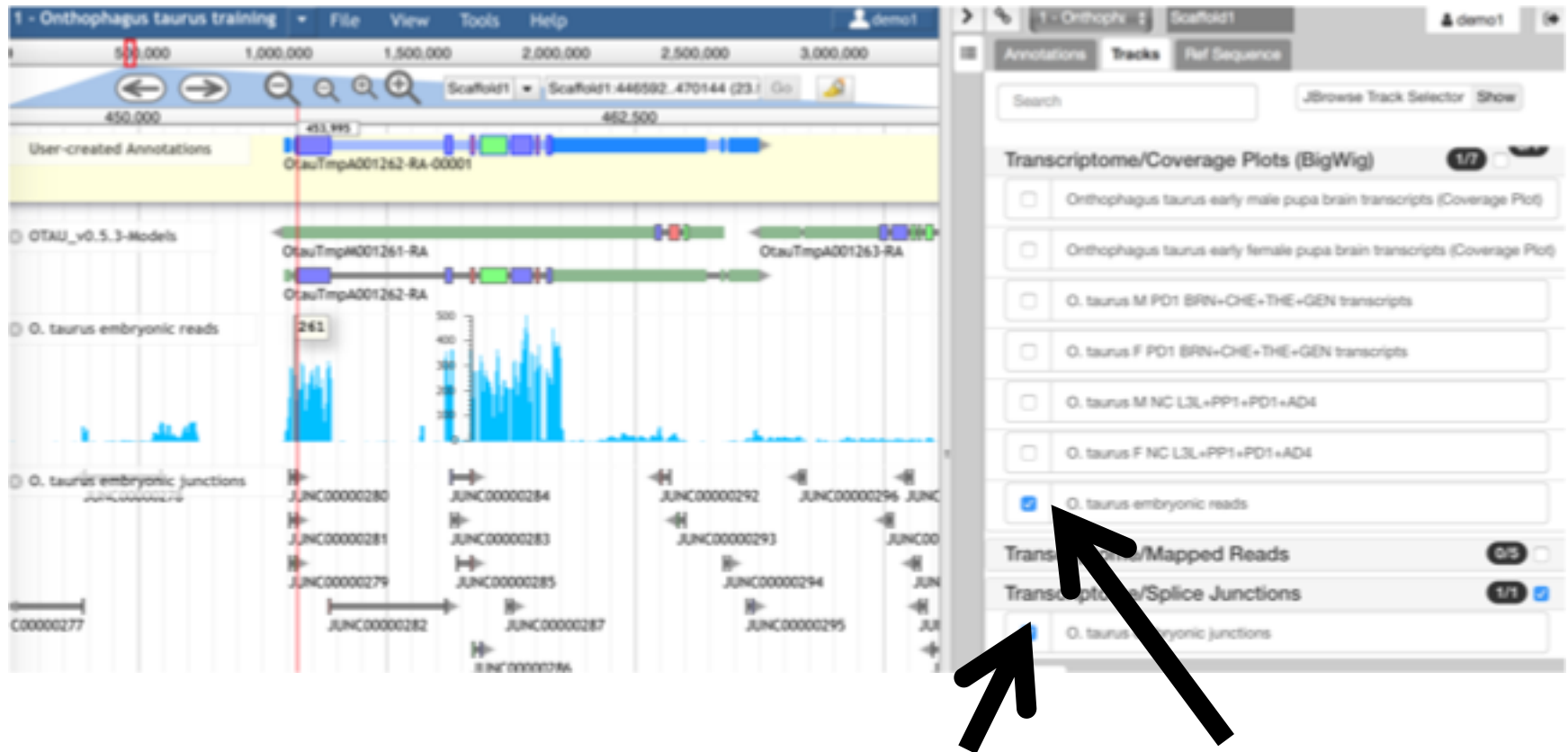


Drag model OtauTimpM001261-RA to User-created Annotations track

# Modify *O. taurus* model sequence in Apollo

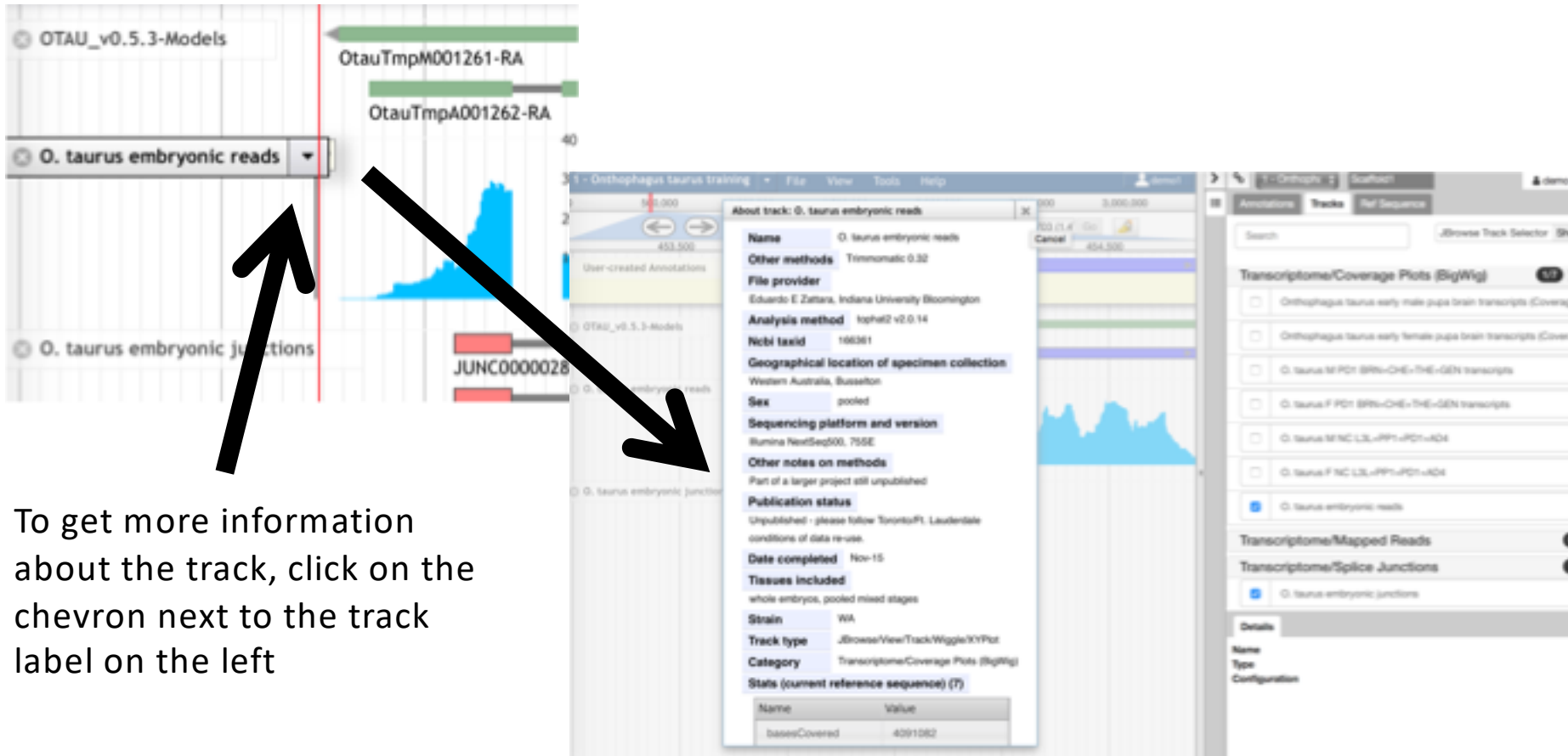
- Questions:
  - What evidence do you choose to check the integrity of the model?
  - Do you need additional evidence?
  - How do you evaluate whether the protein sequence is as complete as it can be?
  - Should you add/modify UTRs?
  - Should you add isoforms?

# View available evidence



Pick Transcriptome coverage plot and Splice Junctions.  
Looks okay so far; let's do an alignment to see if/where  
the protein sequence can be improved

# View available evidence



The screenshot displays a genomic browser interface. On the left, a track labeled 'O. taurus embryonic reads' has a downward-pointing chevron next to its label. A red vertical line is positioned over a track. Two large black arrows point from the text below to the chevron and label of the 'O. taurus embryonic reads' track. The main panel shows a genomic track with a blue coverage plot and a red vertical line. A pop-up window titled 'About track: O. taurus embryonic reads' is open, displaying metadata for the track. The right panel shows a list of tracks, including 'Transcriptome/Coverage Plots (BigWig)' and 'Transcriptome/Splice Junctions'.

To get more information about the track, click on the chevron next to the track label on the left

Name	Value
basesCovered	4291282

# ClustalO alignment to check completeness of protein sequence

<https://i5k.nal.usda.gov/webapp/clustal/>

```
sp|P35220|CTNA_DROME      KFDQKVGAAGVGLSNNSNKKVDENDFIDASHLVYDGVNEIRRAVLNHSSEDLDTOTFE
TC004609                  KFSQKVQVAVQALSSVPTKEVDENDFIDASHLVYDGVNEIRRAVLNRADEDLDPE-DVE
OtauTnpM001261-RA        KFGQKVAVAVSALSSNPAKVDENDFIDASHLVYDGVNEIRRAVLNRADEDLDPE-DVE
                          **,***_**_***_**_*:*****:****_:**_*

sp|P35220|CTNA_DROME      PVEDLTLETSSSSAHTGDQTVDEYDPDISGICTAREAMKKMTEEDFQKIAQQVELFRREK
TC004609                  LDENYTLETS*SSAHTGEHGVDEYDPDISGITTAREAMGKMPEDFQKILQQVEYFRSEK
OtauTnpM001261-RA        LGENTPYDNRS*SSAHTGEHGVDEYPEISGITTAREAMGKMPEDFQKILQQVEFFRSEK
                          *:  :.***:*****: *****:***** ** ***** ** **

sp|P35220|CTNA_DROME      LTFDSEVAKWDDTGNDIIFLAKHMCIMHEMTDFTGRGPLKTTMDVINAAKKISEAGTK
TC004609                  LKFDREVAKWDDTGNDIIVLAKHMCIMHEMTDFTGRGPLKTTMDVINAAKKISEAGTK
OtauTnpM001261-RA        LIFDREVAKWDDTGNDIIVLAKHMCIMHEMTDFTGRGPLKTTMDVINAAKKISEYGTK
                          * ** *****:*****:*****:***** **

sp|P35220|CTNA_DROME      LDKLTNQLAEQCPESSTKKDLLAYLQRIALYCHQIQITSKVKADVQNISGELIVSGL--D
TC004609                  LDKLTNQLAEQCPESSTKKDLLAYLQRIALYCHQMNTITSKVKADVQNISGELIVSGL--D
OtauTnpM001261-RA        LDKLTNQLADQCPESSTKKDLLAYLQRIQLYCHQMNTITSKVKADVQNISGELIVSGVNL
                          *****:***:*****:***** *****:*****:*****:*****

sp|P35220|CTNA_DROME      SATSLIQAANKNLNNAVVLTVKYSYVASTKYFQQTIVSSPIVVKMKAFKKPLVRPEKFE
TC004609                  SATSLIQAANKNLNNAVVLTVKASYVASTKYFQQTIVSSPIVVKMKAFKKPLVRPEKFE
OtauTnpM001261-RA        SATSLIQAANKNLNNAVVLTVKASYVASTKYFQHTIVSSPIVVKMKAFKKPLVRPEKFE
                          *****:*****:*****:***** *:***:*****:*****:*****

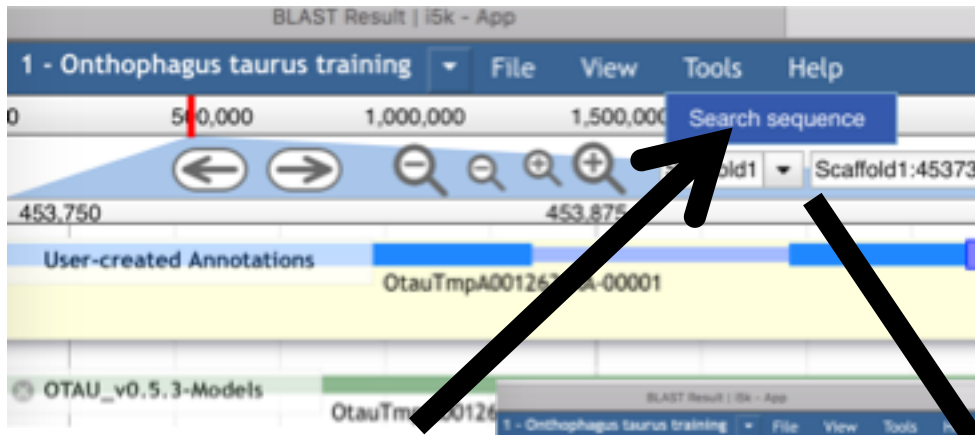
sp|P35220|CTNA_DROME      EVRAKVRXGSKKQVQNPPIHALSEFQSPADAV
TC004609                  EVRAKVRXGSKKQVQNPPIHALSEFQSPTESI
OtauTnpM001261-RA        EARAkVRXGAQKKVQNPPIHALSEFQSPTESV
                          *,*****:*****:*****:*****:*****
```



Alignment looks pretty good – just 2 residues might need to be fixed

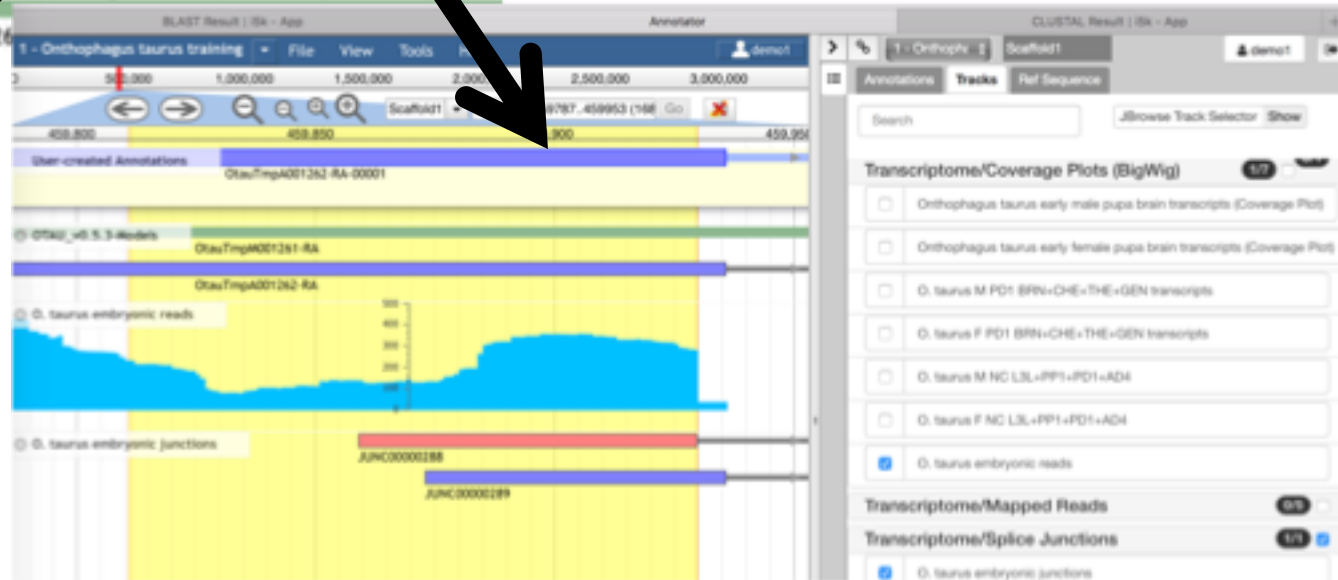
uncolorful To hmmsearch

# Use BLAT to locate problem area

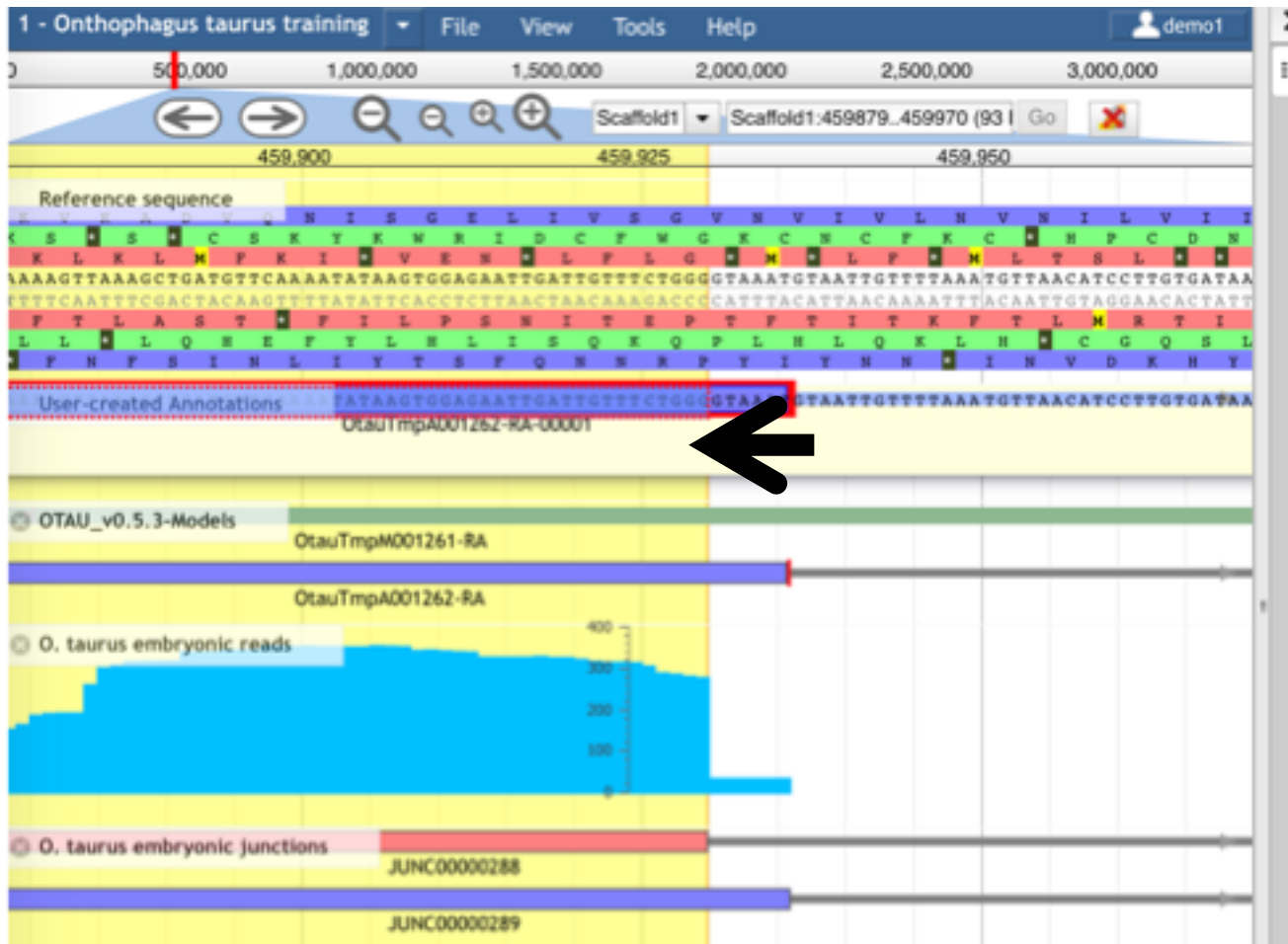


Result is highlighted in yellow

Find 'Search sequence'  
under Tools menu



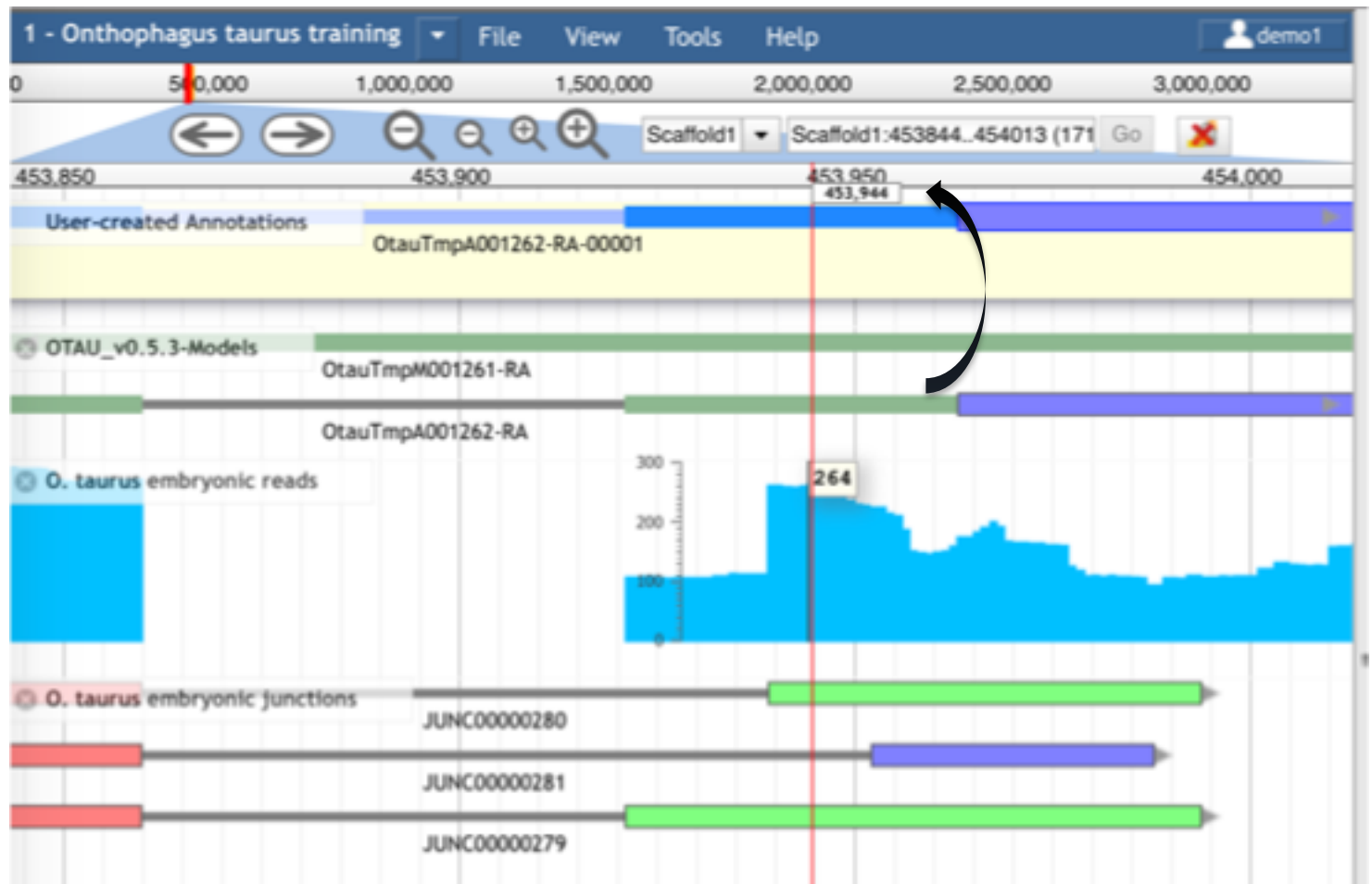
# Adjust exon boundary



Drag exon boundary to match majority of RNA-Seq evidence

# Possible isoform in 5' UTR

Drag original model to UCA again to create a new isoform; or right-click model in UCA and select “Duplicate”



# Evaluate new protein sequence

- Blast modified OtauTmpM001262-RA sequence to NCBI's nr database
  - Make sure it doesn't match a potential contaminant
  - Get an idea whether you have the right sequence
  - Blastp home:
    - [https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE\\_TYPE=BlastSearch&LINK\\_LOC=blasthome](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)

# Using the Information Editor

The screenshot shows the 'Information Editor' window for 'Onthophagus taurus training'. It is divided into two main sections: 'gene' on the left and 'mRNA' on the right. Each section has a 'Select mRNA' dropdown at the top. Below this, there are fields for 'Name', 'Symbol', 'Description', 'Created', and 'Last modified'. Further down are sections for 'DBXRefs' (with 'DB' and 'Accession' columns), 'Replaced Model' (with 'Tag' and 'Value' columns), 'PubMed IDs', and 'Gene Ontology IDs'. Each of these sections has 'Add' and 'Delete' buttons. The 'mRNA' section is currently selected, and the 'Name' field contains 'alpha catenin'.

Use the  
mRNA/transcript side  
of the IE

Review our naming guidelines  
before naming:

<https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>

We're in the process of  
deprecating the 'Replaced  
Model' field – no need to use

# Using the Information Editor

- Select the model in Apollo, then right-click, and select 'Edit Information' from the drop-down menu
  - Use the 'mRNA' section
  - **Please review our naming guidelines:**
    - <https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>
    - If a naming convention exists, use it (e.g. for gene families)
    - Use name from an orthologous protein if you are sure that your gene model is orthologous.
    - Document your justification for the name in the Comments field (e.g. "88% sequence similarity via blastp to D. melanogaster pepck P20007")
    - If you create a new name, it should be unique and attributed to all orthologs (as far as possible)
    - Comments – Document what changes you performed, and your justification for the name. These notes will be visible in the OGS, so make sure that others understand them

# Checklist for accuracy and integrity

- Check start, stop and exon boundaries (splice sites)
  - Try to fix non-canonical splice sites if possible
- Check if you can annotate UTRs (e.g. using RNA-Seq data)
- Check for gaps in the genome
- If you change the genome sequence, add a justification comment to the corresponding gene model
- Use BLAST or a multiple sequence aligner
  - To look at completeness of model
  - To verify the appropriateness of the gene name
- In the Information editor **mRNA** field
  - Update the Name if appropriate
  - Add comments that describe
    - your evidence for the annotation
    - Modifications that you made to the gene model

cf. <https://www.slideshare.net/MonicaMunozTorres/editing-functionality-apollo-workshop>

# What happens to my annotation when I'm done?

- This depends on the genome project that you're working on.
- If the genome coordinator has asked us to generate an OGS (Official Gene Set), we will do so
  - We are still working on this process, so if you ask us to do this, 1) it will take some time, and 2) we will probably ask you for co-authorship if you publish a paper on the OGS.
  - You can also try out the process yourself: <https://github.com/NAI-i5K/GFE3toolkit/>
  - We are working on a pipeline to submit Official Gene Sets to GenBank, where they will be archived/accessioned
- Otherwise, don't assume that your annotation will be archived.
  - If you need it to be, get in touch with us and we'll figure out what to do.
- Get in touch with us and the genome project coordinator if you're not sure about the status of a genome project.
- <https://i5k.nal.usda.gov/data-management-policy>

# Thank you!

## The NAL Team

- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Gary Moore
- Susan McCarthy

## I5k Workspace alumni

- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

- i5k Coordinating Committee
- i5k Pilot Project
- Apollo & JBrowse Development Teams
- GMOD/Tripal community
- All of our users and contributors!

## Contact us:

- <https://i5k.nal.usda.gov/contact>
- [i5k@ars.usda.gov](mailto:i5k@ars.usda.gov)
- [Monica.Poelchau@ars.usda.gov](mailto:Monica.Poelchau@ars.usda.gov)
- [Christopher.Childers@ars.usda.gov](mailto:Christopher.Childers@ars.usda.gov)

PLEASE fill out the post-workshop survey!

<https://tinyurl.com/ybppr8pq>

## Part 2: Hands-on exercises

- We've set up separate Apollo sites for each of you, containing a subset of the *O. taurus* genome assembly
- Sign up for one of the Apollo sites here:
  - <https://tinyurl.com/ycg5bmtk>
- The i5k Workspace has just recently started moving to Apollo2, so you might notice some issues – feel free to let us know

# Part 2: Hands-on exercises

- Two *Onthophagus taurus* examples for you to work on.
  - Medium difficulty: Phosphoenolpyruvate carboxykinase (pepck)
  - Hard: Couch potato (cpo)
- Use the resources described in this tutorial to:
  - Find appropriate reference genes
  - Identify likely homologs in *O. taurus*
    - Training Blast -> Apollo
  - Improve the *O. taurus* structural annotations, if necessary
  - Add functional annotation

# Part 2: Hands-on exercises

- Important URLs (text file with these is in the online course folder):
  - URLs to UniProt pages for pepck and cpo
  - Fasta file for alpha-catenin
  - Training Blast site:  
<https://i5k.nal.usda.gov/training/webapp/blast>
  - Apollo (if you want to access without Blast):  
<https://apollo.nal.usda.gov/apollo/jbrowse/>
  - Uniprot: <http://www.uniprot.org/>
  - OrthoDB: <http://www.orthodb.org/>